Dynamic Sentence Sampling for Efficient Training of Neural Machine Translation

Rui Wang, Masao Utiyama, and Eiichiro Sumita
National Institute of Information and Communications Technology (NICT)
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan
{wangrui, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Traditional Neural machine translation (NMT) involves a fixed training procedure where each sentence is sampled once during each epoch. In reality, some sentences are well-learned during the initial few epochs; however, using this approach, the well-learned sentences would continue to be trained along with those sentences that were not well learned for 10-30 epochs, which results in a wastage of time. Here, we propose an efficient method to dynamically sample the sentences in order to accelerate the NMT training. In this approach, a weight is assigned to each sentence based on the measured difference between the training costs of two iterations. Further, in each epoch, a certain percentage of sentences are dynamically sampled according to their weights. Empirical results based on the NIST Chinese-to-English and the WMT English-to-German tasks show that the proposed method can significantly accelerate the NMT training and improve the NMT performance.

1 Introduction

Recently neural machine translation (NMT) has been prominently used to perform various translation tasks (Luong and Manning, 2015; Bojar et al., 2017). However, NMT is much more time-consuming than traditional phrase-based statistical machine translation (PBSMT) due to its deep neural network structure. To improve the efficiency of NMT training, most of the studies focus on reducing the number of parameters in the model (See et al., 2016; Crego et al., 2016; Hubara et al., 2016) and implementing parallelism in the data or in the model (Wu et al., 2016; Kalchbrenner et al., 2016; Gehring et al., 2017; Vaswani et al., 2017).

Although these technologies have been adopted, deep networks have to be improved to achieve state-of-the-art performance in order to handle very large datasets and several training iterations. Therefore, some researchers have proposed to accelerate the NMT training by resampling a smaller subset of the data that makes a relatively high contribution, to improve the training efficiency of NMT. Specifically, Kocmi and Bojar (2017) empirically investigated curriculum learning based on the sentence length and word rank. Wang et al. (2017a) proposed a static sentence-selection method for domain adaptation using the internal sentence embedding of NMT. They also proposed a sentence weighting method with dynamic weight adjustment (Wang et al., 2017b). Wees et al. (2017) used domain-based cross-entropy as a criterion to gradually fine-tune the NMT training in a dynamical manner. All of these criteria (Wang et al., 2017a,b; Wees et al., 2017) are calculated before performing the NMT training based on the domain information and are fixed while performing the complete procedure. Zhang et al. (2017) adopted the sentence-level training cost as a dynamic criterion to gradually fine-tune the NMT training. This approach was developed based on the idea that the training cost is a useful measure to determine the translation quality of a sentence. However, some of the sentences that can be potentially improved by training may be deleted using this method. In addition, all of the above works primarily focused on NMT translation performance, instead of training efficiency.

In this study, we propose a method of dynamic sentence sampling (DSS) to improve the NMT training efficiency. First, the differences between

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 298–304
Melbourne, Australia, July 15 - 20, 2018.

the training costs of two iterations, which is a measure of whether the translation quality of a sentence can be potentially improved, is measured to be the criterion. We further proposed two sentence resampling strategies, i.e., weighted sampling and review mechanism to help NMT focus on the not well-learned sentences as well as remember the knowledge from the well-learned sentences.

The remainder of this paper is organized as follows. In Section 2, we introduce the dynamic sentence sampling method. Experiments are described and analyzed in Section 3. We discussed some other effects of the proposed methods in Section 4. We conclude our paper in the last section.

2 Dynamic Sentence Sampling (DSS)

2.1 NMT Background

An attention-based NMT system uses a bidirectional RNN as an encoder and a decoder that emulates the search through a source sentence during the decoding process (Bahdanau et al., 2015; Luong et al., 2015). The training objective function to be minimized can be formulated as:

$$J = \sum_{\langle x,y\rangle \in D} -\log P(y|x,\boldsymbol{\theta}), \quad (1)$$

where $\langle x, y\rangle$ is the parallel sentence pair from the training corpus D, $P(y|x)$ is the translation probability, and $\boldsymbol{\theta}$ is the neural network parameters.

2.2 Criteria

The key to perform sentence sampling is to measure the criteria. As we know, the NMT system continually alters throughout the training procedure. However, most of the criteria described in the introduction remain constant during the NMT training process. Zhang et al. (2017) adopted the sentence-level training cost to be a dynamic criterion; further, the training cost of a sentence pair $\langle x, y\rangle$ during the ith iteration can be calculated as:

$$cost^i_{\langle x,y\rangle} = -\log P(y|x,\boldsymbol{\theta}). \quad (2)$$

Directly adopting training cost as the criterion to select the top-ranked sentences that represent the largest training costs has two drawbacks: 1) The translation qualities of sentences with small training costs may be further improved during the succeeding epochs. 2) If the training corpus become smaller after each iteration, the knowledge associated with the removed sentences may be lost over the course of the NMT process.

Therefore, we adopt the ratio of differences (dif) between training costs of two training iterations to be the criterion,

$$dif^i_{\langle x,y\rangle} = \frac{cost^{i-1}_{\langle x,y\rangle} - cost^i_{\langle x,y\rangle}}{cost^{i-1}_{\langle x,y\rangle}}. \quad (3)$$

It should be noted that some of $dif_{\langle x,y\rangle}$ are negative. That is, the costs of some sentence pairs even increase after one epoch training. Therefore, the difference is normalized into [0, 1] as the final criterion:

$$criterion^i_{\langle x,y\rangle} = \frac{dif^i_{\langle x,y\rangle} - min(dif^i)}{max(dif^i) - min(dif^i)}. \quad (4)$$

This criterion indicates the likelihood of a sentence to be further improved in the next iteration; low values indicate that the training cost of a sentence is unlikely to change and that it would not significantly contribute to the NMT training even if the sentence was trained further.

2.3 Dynamic Sampling

As we know, the NMT performance improves significantly during the initial several epochs and less significantly thereafter. This is partially because that some of the sentences have been learned sufficiently (i.e., low $criterion^i_{\langle x,y\rangle}$ values). However, they are kept further training with the ones which have not been learned enough (i.e., high $criterion^i_{\langle x,y\rangle}$ values). Therefore, in this approach, these sentences are deleted for the subsequent iterations. To ensure that knowledge from the deleted sentences is retained, we propose two mechanisms for dynamic sampling, which are described in the succeeding sections.

2.3.1 Weighted Sampling (WS)

We assign a normalized weight to each sentence according to the criterion that can be given as:

$$weight^i_{\langle x,y\rangle} = \frac{criterion^i_{\langle x,y\rangle}}{\sum_{\langle x,y\rangle \in D} criterion^i_{\langle x,y\rangle}}. \quad (5)$$

Further, weighted sampling without any replacement was used to select a small subset,

such as 80%[1] of the entire corpus, as the corpus D_{ws}^{i+1} to perform the subsequent iteration. The updated objective function using weighted sampling J_{ws} can be formulated as follows:

$$J_{ws} = \sum_{\langle x,y\rangle \in D_{ws}} -\log P(y|x,\boldsymbol{\theta}). \quad (6)$$

Thus only 80% of the entire corpus is used to perform the NMT training during each iteration (for the first two iteration, all of the sentences should be sampled). Because the criterion continually changes, the sentence selection procedure also changes during the NMT training. Those that are not selected in an epoch still have a chance to be selected in the subsequent epoch[2].

2.3.2 Review Mechanism (RM)

We further propose an alternate sentence sampling mechanism. After performing an iteration during training, 80% of the top-ranked sentences are selected to act as the training data for the subsequent iteration. Each sentence that is not selected is classified into the low-criterion group D_{low} and does not have a chance to be sampled again. In this case, the D_{low} will become larger and larger, and D_{high} will becomes smaller and smaller. To prevent the loss of the knowledge that was obtained from the D_{low} group during NMT, a small percentage λ, such as 10%, of the D_{low} group is sampled as the knowledge to be reviewed. The updated NMT objective function is formalized as follows,

$$J_{rm} = \sum_{\langle x,y\rangle \in D_{high}} -\log P(y|x,\boldsymbol{\theta}) + \sum_{\langle x,y\rangle \in \lambda D_{low}} -\log P(y|x,\boldsymbol{\theta}). \quad (7)$$

3 Experiments

3.1 Datasets

The proposed methods were applied to perform 1) the NIST Chinese (ZH) to English (EN) translation task that contained a training dataset of 1.42 million bilingual sentence pairs from LDC corpora[3]. The NIST02 and NIST03-08 datasets were used as the development and test datasets, respectively. 2) the WMT English to German (DE) translation task for which 4.43 million bilingual sentence pairs from the WMT-14 dataset[4] was used as the training data. The newstest2012 and newstest2013-2015 datasets were used as development and test datasets, respectively.

3.2 Baselines and Settings

Beside the PBSMT (Koehn et al., 2007) and vanilla NMT, three typical existing approaches described in the introduction were empirically compared: 1) Curriculum learning using the source sentence length as the criterion (Kocmi and Bojar, 2017). 2) Gradual fine-tuning using language model-based cross-entropy (Wees et al., 2017)[5]. 3) NMT boosting method by eliminating 20% of the training data with the lowest training cost after performing every iteration (Zhang et al., 2017).

For the proposed DSS method, we adopted one epoch as one iteration for the EN-DE task and three epochs as one iteration for the ZH-EN task, because the corpus size of the EN-DE task is approximately three times larger than that of the ZH-EN task.

3.3 NMT Systems

The proposed method was implemented in Nematus (Sennrich et al., 2017) with the following default settings: the word embedding dimension was 620, the size of each hidden layer was 1,000, the batch size was 80, the maximum sequence length was 50, and the beam size for the decoding was 10. A 30K-word vocabulary was created and data was shuffled before each epoch. Training was conducted on a single Tesla P100 GPU using default dropout and the ADADELTA optimizer (Zeiler, 2012) with default learning rate 0.0001. All of the systems were trained for 500K batches which took approximately 7 days.

[1] Zhang et al. (2017) adopted 80% as the selection threshold and we follow their settings for fair comparison. Due to limited space, we will empirically investigate the effect of the thresholds as our future work.

[2] For those 20% sentences who are not selected, their $criterion_{\langle x,y\rangle}^{i+1} = criterion_{\langle x,y\rangle}^{i}$.

[3] LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08, and LDC2005T06.

[4] https://nlp.stanford.edu/projects/nmt/data/wmt14.en-de/

[5] Wees et al. (2017) also proposed a weighted sampling method; however, its performance was worse than that of the gradual fine-tuning. The method originally adopted by Wees et al. was based on the cross-entropy differences between two domains. Because no domain information is available for this task; the development data was used as the in-domain data by that method. In the method proposed in this study, the development data is not required.

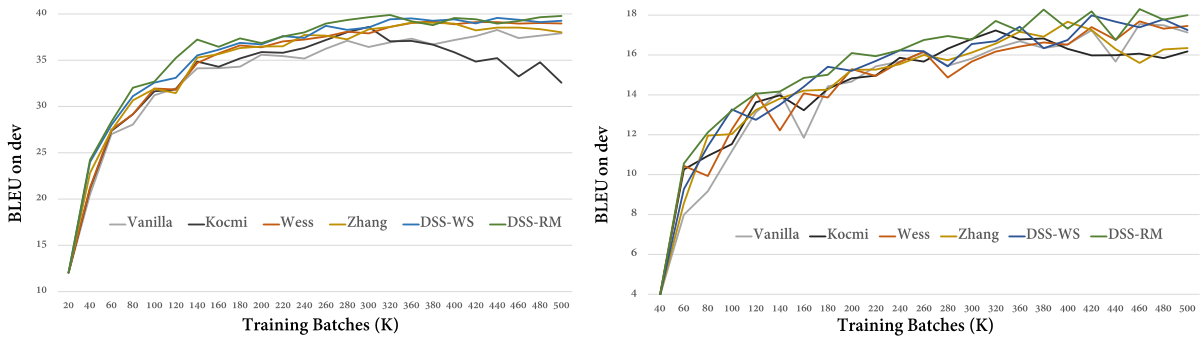

Figure 1: Learning curves. Left: NIST ZH-to-EN; Right EN-to-DE.

Table 1: Results from the NIST ZH-to-EN translation task.

Methods	Dev (NIST02)	NIST03	NIST04	NIST05	NIST06	NIST08	Test (all)
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Vanilla NMT	38.48	37.53	39.95	35.24	33.86	27.23	35.08
Random Sampling	38.35	36.45	40.01	34.27	33.70	26.37	34.62
Kocmi and Bojar (2017)	38.51	37.60	39.87	35.43	33.76	27.37	35.19
Wees et al. (2017)	**39.16**	38.09	40.30	**35.59**	**34.14**	27.46	**35.62**
Zhang et al. (2017)	39.08	**38.27**	**40.37**	35.32	33.57	**27.87**	35.57
DSS-WS	39.54+	39.23++	**40.84+**	**35.98+**	34.91++	28.42+	36.85++
DSS-RM	**39.89++**	**39.90++**	40.60	35.77+	**35.45++**	**29.30++**	**37.33++**

Table 2: Results from the WMT EN-to-DE translation task.

Methods	Dev (newstest2012)	newstest2013	newstest2014	newstest2015	Test (all)
PBSMT	14.89	16.75	15.19	16.84	16.35
Vanilla NMT	17.55	20.92	19.16	20.01	20.06
Random Sampling	17.39	20.32	18.36	20.30	19.61
Kocmi and Bojar (2017)	17.63	20.63	19.21	**20.47**	20.18
Wees et al. (2017)	**17.69**	**20.81**	19.21	20.24	20.19
Zhang et al. (2017)	17.67	20.80	**19.37**	20.42	**20.30**
DSS-WS	17.99	21.11	19.89+	**21.20+**	20.96+
DSS-RM	**18.34+**	**21.76++**	**20.04++**	21.02+	**21.22++**

Note: The translation performance was measured using the case-insensitive BLEU (Papineni et al., 2002) scores. Marks after the scores indicate that the proposed methods significantly (Koehn, 2004) outperformed the existing optimal baselines in bold ("++"denotes better at a significance level of $\alpha = 0.01$, whereas "+"denotes better at a significance level of $\alpha = 0.05$.).

3.4 Results and Analyses

3.4.1 Training Efficiency

The learning curve is depicted in Figure 1.

1) The BLEU score (ZH-EN as example) of vanilla NMT increased from 0 to 35 using approximately 200K training batches. Further, the BLEU increased from 35 to 38 using around 200K additional training batches. This is consistent with our hypothesis that the improvement in NMT shows decreasing significance as the training progresses.

2) For the baselines, the method developed by Kocmi and Bojar (2017) did not provide significant improvement in speed. The method proposed by Wees et al. (2017) and Zhang et al. (2017) slightly accelerated the NMT training.

3) The proposed DSS methods significantly accelerated the NMT training. The BLEU score (ZH-EN as example) reached 35 after using approximately 140K training batches; further, the BLEU score reached 38 after using approximately additional 120K training batches. This may be caused due to the fact that the amount of well-learned became larger and larger as the training kept going. If these sentences were continually trained, the performance would not increase significantly. In comparison, DSS methods eliminated these well-learned sentences; therefore, the performance kept improving significantly until all of the sentences become well-learned.

4) The performances of Kocmi and Bojar

(2017) and Zhang et al. (2017) decreased significantly after reaching the highest BLEU. This is consistent with the hypothesis that NMT may forget the learned knowledge by directly removing corresponding sentences. In comparison, the performances of the proposed DSS methods did not decrease significantly, because the removed sentences still have chances to be sampled.

3.4.2 Translation Performance

For fair comparison, we evaluated the best performed (on dev data) model during 500K training batches on the test data. The results are shown in Tables 1 and 2.

1) The methods proposed by Wees et al. (2017) and Zhang et al. (2017) slightly improved performances. On Test(all), the proposed DSS methods significantly improved the BLEU score by approximately 1.2∼2.2 as compared to the vanilla NMT and by 0.9∼1.7 to the best performing baselines. As the well-learned sentences increases during NMT training, it did not only slow down NMT training, but also prevent NMT from learning knowledge from the sentences which were not well learned and cause the improvement stagnate.

2) Within the DSS methods, the review mechanism appears to be a slightly better mechanism than weighted sampling. This indicates that the review mechanism retained the learned knowledge in a better manner than the learned knowledge of the weighted sampling.

4 Discussions

During the response period, the comments and suggestions of reviewers inspired us a lot. Due to the limited time and space, we briefly discussed these suggestions in this paper. We will show the empirical results in our future work.

4.1 Effect on Extreme Large Data

For the large corpus, we have tested the WMT EN-FR task, which containing approximately 12M sentences. The NMT trained from large-scale corpus still gained slight BLEU improvement after several-epoch training. After 6 epochs training (1M batches), the proposed dynamic sentence sampling method outperformed the baseline by approximately 0.6 BLEU.

For the web-scale corpora which may be converged within one epoch, in our opinion, if a sentence pair is not well-learned enough, it is necessary to learn it once more. To accelerate this judging processing, we can adopt the sentence similarities between the untrained sentence with small-sized trained sentences as the criteria for sentence sampling.

4.2 Effect on Long-time Training

Similarly, for the WMT EN-DE and NIST ZH-EN, if we keep training for more than 1M batches which takes 2-3 weeks, the BLEU would increase by 1.0-1.5 and differences between baseline and the proposed method would slightly decrease by 0.5-0.7 BLEU. Because 7-10 days is a reasonable time for NMT training, we reported 500K batches training results in this paper.

4.3 Effect on Noisy Data

We added 20% noisy data, which is wrongly aligned, to the NIST ZH-EN corpus. Empirical result shows that the training cost of these noise data did not decrease significantly and even increase sometimes during the training processing. After the first-time time dynamic sampling training by the proposed method, the noise data ratio decreased from 20% to 13%. After the second-time dynamic sampling training, the noise data ratio decreased from 13% to 7%. This indicates that the proposed method can also detect the noisy data.

5 Conclusion

In this study, the sentences for which training costs of two iterations do not show any significant variation are defined as well-learned sentences. Using a dynamic sentence sampling method, these well-learned sentences are assigned a lower probability of being sampled during the subsequent epoch. The empirical results illustrated that the proposed method can significantly accelerate the NMT training and improve the NMT performances.

Acknowledgments

Thanks a lot for the helpful discussions of Kehai Chen, Zhisong Zhang, and three anonymous reviewers. This work is partially supported by the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of MIC, Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, San Diego.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Quantized neural networks: Training neural networks with low precision weights and activations. *CoRR*, abs/1609.07061.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *CoRR*, abs/1610.10099.

Tom Kocmi and Ondrej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. *CoRR*, abs/1707.09533.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. *CoRR*, abs/1606.09274.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1483–1489, Copenhagen, Denmark.

Marlies Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1411–1421, Copenhagen, Denmark.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Dakun Zhang, Jungi Kim, Josep Crego, and Jean Senellart. 2017. Boosting neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 271–276, Taipei, Taiwan.